Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Titulo de Tesis

Tesis de Licenciatura en Ciencias de la Computación

Damián Eliel Aleman

Director: Juan Manuel Pérez y Santiago Kalinowski
Codirector: Agustín Gravano
Buenos Aires, 2017

Índice general

1. INTRODUCCIÓN

1.0.1. Trabajo previo en el área

Actualmente, las herramientas para la detección de palabras con contraste léxico en distintas regiones consisten en cuestionarios como el de *Almeida*[AV95]. Estos cuestionarios están integrados por grupos temáticos centrales como la casa, la familia, la enseñanza, el cuerpo humano, etc. Sobre cada grupo temático se les indicaba a las personas entrevistadas un repertorio de palabras por cada noción, para ver si las conocían y con que frecuencia se las usaba. Con este trabajo planteamos cambiar el paradigma y detectar automáticamente las palabras usadas en distintas regiones y sus frecuencias.

Uno de los corpus lingüísticos del español más reconocidos es el *corpes XXI*[Esp], creado por la Real Academia Española con una distribución de 25 millones de formas por cada uno de los años comprendidos en el periodo 2001 a 2012. Sin embargo, dicho corpus tiene dos desventajas importantes: por un lado, la cantidad de palabras de América Latina están subrepresentados ya que el $65,70\,\%$ de las palabras del dataset provienen de textos de España y $34,30\,\%$ de los demás países hispanoparlantes. Por otro lado, uno no dispone de todo el dataset, sino que solamente se pueden hacer las consultas de su página web. Estas consultas están limitadas en cuanto a la cantidad de solicitudes y a las funcionalidades que estas proveen.

Una de las virtudes de hacer un corpus con un método de recolección de textos de forma automática se desprende de la cantidad superadora de longitud del corpus en comparación con métodos manuales como digitalización de textos.

A pesar de haber comenzado hace varias decadas la recolección de textos de la web para realizar corpus, no hay muchos en el idioma español. Uno que se pudo encontrar es el de *Mark Davies*, el cual se utilizó las páginas web para recolectar los textos, con dos billones de palabras en español y divide a las páginas en regiones a través de

Las ventajas de Twitter son claras: da una interfaz pública para obtener tweets de cualquier persona. Además a diferencia de un portal de noticias donde los comentarios suelen estar relacionados con estas, en Twitter son más amplio los tópicos de los comentarios. Por otro lado, Twitter es una tecnología que permite hacer escalabale el trabajo a diferentes paises ya que con una misma interfaz se pueden obtener los textos de cualquier región. En cambio, si se elige un portal de noticias para sacar comentarios de usuarios, deberíamos ver la estructura de cada página para obtener esos datos. Otra ventaja de Twitter es que cada usuario está identificado y se podría llegar a inferir datos como género, edad y ubicación de cada uno. Por último una ventaja sobre otro método de recolección, es que a través de Twitter se puede recolectar textos con una granularidad regional muy variable y obtener información de quienes lo escriben. Existen otras redes sociales que se podrían obtener textos para analizar , pero tienen la desventaja de ser privadas como Facebook o ser acotadas en términos de los temas que se hablan en la misma, un caso de esto es Linkedin.

1.0.2. Twitter

Twitter es un servicio de microblogging creado en el 2006. Los usuarios son variados, desde personas, instituciones gubernamuentales, no gubernamentales y bots(i.e programa que corre tareas automáticamente). Cada usuario puede escribir textos llamados tuits, que tienen una longitud máxima de 140 caracteres. Las relaciones en Twitter no necesitan ser recíprocas, uno puede seguir a una persona, en cuyo caso va a poder leer todos los tuits generado por ella, como también puede ser seguido por una persona. Un tuit puede ser

respondido, como también puede ser retuiteado. El retuit es un mecanismo para diseminar por la red tuits generados por otros usuarios. De esta manera si un usuario A realiza un retuit generado por el usuario B, cualquier seguidor de A también va a recibir ese tuit en su panel, al cual llamaremos t*timeline*. Si bien los tuits que se ven en el timeline son solo aquellos generados por los usuarios que uno sigue, todos los tuits son públicos, es decir que pueden ser accedidos a tavés de búsquedas en la plataforma. Para dar una noción de la cantidad de usaurios en la Argentina,en el 2016 había 11,8 millones de usuarios de Twitter en la Argentina, con 15 millones de personas con smartphones, lo cual el 70 de la gente con smartphones tenía Twitter.

Qué es Twitter Nombrar trabajos donde se utilizo Ventajas, desventajas

1.0.3. Lingüistica computacional

[Baa01] [KG03] [MH11]

2. MATERIALES Y MÉTODO

2.1. Extracción de Datos

Para la recolección de tweets, primero se extrajo una cantidad de usuarios de forma localizada con el fin de obtener todos los tweets de estos. Los usuarios se buscaron por provinicia de modo tal que haya una cantidad aproximadamente equitativa. La búsqueda de los usuarios se hizo de la siguiente manera:

Por cada provincia de la Argentina, se extrajo las ubicaciones de cada uno de sus departamentos , de los partidos de la provincia de Buenos Aires y de las comunas de la Ciudad Autónoma de Buenos Aires. El conjunto de estas forman la subdivisión de segundo orden de la republica Argentina. La lista de departamentos/partidos/comunas fue extraída a partir de los datos publicados del Censo Argentino realizado en el año 2010. Para extraer los tweets se utilizó la librería de *python* llamada *tweepy*. De esta manera se recolectó aproximadamente 2000 usuarios por provincia lo que resume en 46000 usuarios argentinos. Sobre este conjunto de usuarios se buscaron los tweets realizados por estos. Se decidió no tener en cuenta los retweets dado que estos no son escritos por los usuarios sino que son una mera copia de otros tweets.

2.1.1. Búsqueda geolocalizadas

Las búsquedas geolocalizadas de la API de *twitter* primero intentan de buscar tweets cuyas coordenadas sean las buscadas. En caso de no tener éxito, buscará aquellos tweets creados por usuarios que tienen en el campo *location* de su perfil un lugar cuyo geocódigo coincida con el de sus coordenadas. Es decir, si se hace una búsqueda inversa de las coordenadas, devuelve el lugar de su perfil.

Una vez obtenida la lista de ubicaciones, se realizaron búsquedas por cada provincia con centro en las coordenadas de los departamentos de la misma y con un radio de 20 millas. Sobre el resultado de esta búsqueda, únicamente se seleccionaron los usuarios que tienen como campo *location* al menos uno de los nombres de las ciudades de la provincia. Con esta precaución eliminamos los posibles tweets de turistas que escribieron en un lugar pero que no viven allí.

A continuación se muestra un gráfico con las ubicaciones donde se encuentran los usuarios de la muestra de desarrollo:

FALTA GRÁFICO Si bien en este trabajo nos enfocamos en las coordenadas de las localidades dentro de Argentina, basta con cambiar las coordenadas y los nombres de las localidades que tienen que tener los campos *location* para realizar un análisis sobre otros países.

2.2. Datos de desarrollo y de validación

Por cada provincia se tomó a los usuarios de la misma y se los dividió para tener un conjunto de datos de desarrollo y uno de validación. El conjunto de validación fue creado para poder corroborar que los resultados obtenidos por el análisis del conjunto de desarrollo no sean algo intrínseco de esta muestra, sino que se pueden extrapolar a toda la población. La división de los datos se realizó de manera tal que los conjuntos resultantes sean lo más independientes posibles:

Usuarios disjuntos Debido a que ciertos usuarios repiten palabras constantemente, ya sea porque son bots o simplemente porque hablan siempre de los mismos temas, es

adecuado validar los resultados con textos producidos por distintos usuarios. De esta manera se intentó mitigar el ruido generado por estos usuarios particulares.

Fechas disjuntas Al analizar los resultados sobre los textos generados en un tiempo acotado de tiempo, estamos trabajando con una muestra específica que es de esperar que tenga fenómenos particulares debido al momento en que fueron escritos. Por ejemplo, debido a cierto fenómeno climático o en el transcurso de un evento polémico (como un debate presidencial o un torneo deportivo) se pueden obtener tweets con una frecuencia de ciertas palabras muy distinta a la frecuencia de la población. Por esta razón se dividió los tweets producidos por los usuarios de manera tal que sus fechas sean disjuntas.

La división fue de la siguiente manera: Sobre el conjunto de usuarios se dividió en dos de forma aleatoria, obteniendo $Usuarios_1$ y $Usuarios_2$. Luego se buscó la fecha $Fecha_{DIV}$ por la cual había una cantidad equiparable entre el conjunto de tweets producidos por $Usuarios_1$ antes de $Fecha_{DIV}$ y el conjunto de tweets producidos por $Usuarios_2$ despúes de $Fecha_{DIV}$.Es decir:

$$Fecha_{DIV} = \arg\min_{F} \left| \sum_{f=FechaInicial}^{F} tweets(Usuarios_1, f) - \sum_{f=F}^{FechaFinal} tweets(Usuarios_2, f) \right| \tag{2.1}$$

Después de fijar la fecha se dividió al conjunto de tweets producidos por estos usuarios: el conjunto de desarrollo con los tweets producidos antes de $Fecha_{DIV}$ y el conjunto de test producidos posteriormente a esa fecha.

2.3. Tokenización y normalización

En cuanto al análisis del texto surge una primer problemática: ¿qué es una palabra?. En principio podemos definir a una palabra como cualquier secuencia de caracteres delimitados por espacios blancos. Con esta definición 523456 y ? serían palabras. Debido a esto podemos restringir nuestra definición a una secuencia de caracteres alfabéticos. Ahora los ejemplos mencionados anteriormente dejarían de estar dentro de la definición. Sin embargo términos como asdsdafsdf también serían palabras. Para restringir aún más la definición podríamos tener un diccionario como filtro para saber si una secuencia de caracteres dada es una palabra. Si bien esto tendría mucha precisión al momento de filtrar los términos, no seríamos capaces de palabras que existen en un lenguaje pero que no están representadas bajo el diccionario elegido. Es por eso que decidimos tomar a una palabra como una secuencia de caracteres alfabéticos.

Es muy posible que tengamos palabras que no sean interesantes a nivel lingüistico, como errores de tipeo (e.g computadira, escribur), errores ortográficos o nombres propios. Es importante destacar que Twitter tiene caracteres especiales para mencionar a la gente, como el @, o el #(hashtag) utilizado para agrupar mensajes. Estos caracteres aparecen mucho, ya que los usuarios suelen responderse en la red, mencionando los mismos temas (aclarando el hashtag), o respondiendo a otros usuarios. Ya que esos caracteres no son alfabéticos, cualquier término que los utilice no va a ser parte del conjunto de palabras, como tampoco lo serán las direcciones de páginsa web. Decidimos que se filtren estos terminos ya que no tienen interés lingüistico y además agregarían mucho ruido a los datos.

Además de la tokenización del texto, se realizó una normalización sobre él. Todas las letras se convirtieron a letra minúscula y las palabras con más de tres letras iguales de forma consecutiva se redujeron para que solo tengan tres repeticiones. De esta forma, el término *padreeeee* y *padreeee* fueron reducidos a una única unidad léxica (*padreee*). Esto se realizó con la librería *TweetTokenizer de NLTK*. Se descartó la idea de filtrar las palabras que no estuvieran en un diccionario ya que si bien hubiera eliminado mucho ruido, también nos hubiera filtrado palabras de interés. Este es el caso de los neologismos, o las palabras que si bien se utilizan hace mucho tiempo no están en los diccionarios actuales.

2.4. Caracterización de la muestra

Para tener una noción más completa de la muestra, presentamos la siguiente tabla que indica las cantidades de palabras y tweets por provincia.

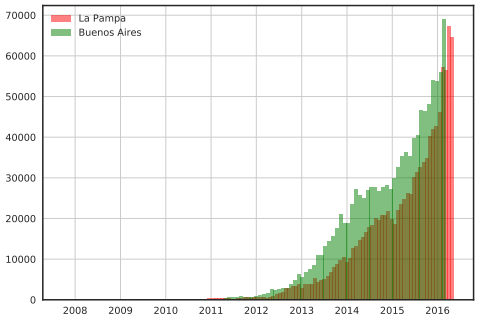
Provincia	#Palabras Distintas	#Usuarios	#Tweets	#Total Palabras
Buenos Aires	191919	920	1125042	8974372
Catamarca	173104	957	1057019	8161309
Chaco	169476	964	976943	7605991
Chubut	182592	954	1023373	8884745
Córdoba	207307	987	1224266	10075932
Corrientes	183292	939	1044951	8426940
Entre Ríos	188679	969	1193693	9462986
Formosa	169254	903	923352	7184382
Jujuy	171064	971	678004	5951778
La Pampa	186593	935	1085757	8996318
La Rioja	186041	946	704044	6757277
Mendoza	193708	945	1099717	9402399
Misiones	168400	972	984218	7790197
Neuquén	188038	927	1111201	9021449
Río Negro	194383	965	1215361	9991831
Salta	188402	884	830916	7506652
San Juan	183546	926	1002322	8377792
San Luis	164185	896	1006464	8327093
Santa Cruz	174089	935	876621	7432923
Santa Fe	201879	937	1019620	8862328
Santiago del Estero	166540	887	944109	7355729
Tierra del Fuego	197273	964	976426	8559218
Tucumán	195643	962	1093874	9238526

Tab. 2.1: Cantidades del conjunto de datos de desarrollo

2.4.1. Cantidad de palabras

2.4.2. Tweets a lo largo del tiempo

Los tweets recolectados para el conjunto de datos de desarrollo tienen una particularidad: a médida que pasan los años hubo mayor cantidad de tweets durante un año. Esto se refleja en los gráficos que presentamos a continuación.



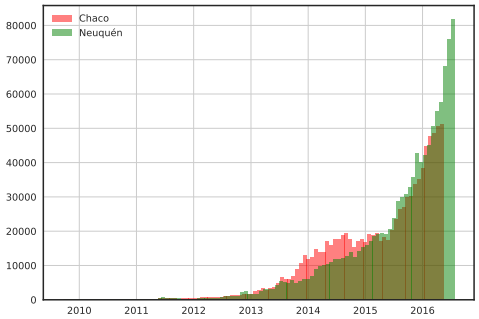


Fig. 2.1:

Fig. 2.2:

Fig. 2.3: En la figura 2.1 se presenta un histograma donde se muestra la cantidad de tweets que se hiceron por intervalo de tiempo en la provincias La Pampa y Buenos Aires. En la figura 2.2, se presenta el gráfico para Chaco y Neuquén.

2.5. Búsqueda de contrastes

Diremos que una palabra tiene un contraste cuando esta tiene un uso con diferencias significativas en distintas regiones. En este trabajo nos propusimos crear un listado con palabras con contrastes que tengan importancia a nivel lingüistico. En este sentido, los nombres de personas, lugares u organizaciones no fueron considerados de interés a pesar de tener contrastes en su uso. Este listado fue ordenado por una métrica que capte en un único valor el nivel contrastivo. De esta manera, se seleccionó un subconjunto de palabras, de acuerdo a la métrica, el cual fue analizado manualmente en otros textos por la Academia Argentina de Letras.

El primer acercamiento para ver el contraste de las palabras lo realizamos comparando las frecuencias de las palabras en cada par de prinvincias de la Argentina. Para esto calculamos el cociente entre la frecuencia máxima de la palabra en las dos provincias sobre la frecuencia mínima, al que llamaremos *maxDif*. En caso de que en una de las dos provincias no se haya recolectado tweets con esa palabra, se tomaba como frecuencia mínima a la frecuencia mínima distinta de 0 de todas las palabras generadas en esa provincia. Así se evitó la división por cero. De esa manera se ordenó el listado de cada par de provincias teniendo en cuenta la división de frecuencias. Sin embargo, este método imposibilitaba el trabajo manual para la Academia Argentina de Letras que debía mirar estos listados y hacer un análisis más exhaustivo sobre las palabras con mayor diferencia de frecuencias, debido a que había $\binom{23}{2} = 253$ listados (o equivalentemente 253 columnas en un mismo listado) a analizar. Además la métrica solo permitía saber si había un contraste entre dos provincias, pero no se podía tener en cuenta la frecuencia de la palabra en el resto de las provincias. En consecuencia las palabras se encontraban repetidas en los distintos listados y con diferentes valores de *maxDif*, lo cual hacía muy díficil poder identificar en que regiones había una diferencia significativa de frecuencias.

Debido a esto decidimos realizar un nuevo enfoque para encontrar las palabras con alta contrastividad en las distintas regiones, de manera que una métrica pueda reflejar el nivel de contrastividad de la palabra en un único valor. De este modo, nos enfocamos a analizar el contraste de frecuencias de palabras sobre las provincias a través de una métrica

superadora.

2.5.1. Métricas para medir el contraste en la frecuencia de las palabras

Dado que se quiere encontrar las palabras con contrastes significativos en distintas regiones se propone generar una métrica basada en la cantidad de información para poder realizar esta tarea.

Una médida que se puede usar para comparar las frecuencias de las palabras en las difentes regiones del país puede ser la entropía definida por Shannon5.0.1, debido a que podemos tener un valor que informe que tan uniforme es la distribución de las frecuencias de cada palabra. Sin embargo, la entropía como única médida tiene sus desventajas. Principalmente, una palabra con una sola ocurrencia en una provincia y ninguna en las demás, tiene la entropía mínima. A pesar de que nos interesan las palabras con un contraste significativo entre regiones, dentro de ellas eligiremos las que tienne mayor cantidad de ocurrencias. Es por esto que elaboramos otra métrica que tenga en cuenta la entropía, pero que no sea la única variable a tener en cuenta.

2.5.2. Valor de información

La métrica que utilizamos para ordenar los listados de palabras y detectar cuales son las que tienen altos contrastes en su uso en distintas regiones fue inspirada sobre el trabajo de Zanette y Montemurro [MZ10]. Ellos a diferencia de Shannon estudiaron una relación entre una médida de la información y su función semántica en el lenguaje. A continuación detallamos el procedimiento para calcular lo que ellos llamaron el valor de la información:

Dado un texto dividido en P partes iguales, se calcula la entropía $H(w)$ sobre el vector de cantidad de ocurrencias en cada una de las P ventanas. Luego se define $\widehat{H(w)}$ como la entropía de una permutación aleatoria del texto y promediada por todos las posibles realizaciones de la permutación de él.

Es decir, se distribuyen uniformemente las palabras en P partes y se calcula la entropía como se hizo con el texto original. Es de esperar que en la mayoría de casos la entropía del texto permutado sea mayor que la médida en el calculo original. Esto se debe a que las palabras se distribuyen de forma más uniforme en las distintas partes. Finalmente, definen al valor de la información como $I(w) = p(w)(\widehat{\eta(w)} - \eta(w))$, con $p(w)$ la frecuencia total de la palabra en el texto. De esta manera se le da más importancia a las palabras que son más frecuentes y a las palabras que tienen una baja entropía, ya que en estas el término de la diferencia es más grande. Este estudio se hizo sobre tres textos, *Análisis de la mente*, *Moby Dick* y *El origen de las especies* de Charles Darwin. En los tres libros las palabras con mayor valor de la información están altamente relacionadas con los temas principales.

Si bien esta métrica tiene en cuenta la frecuencia de las palabras además de la entropía, el texto en Twitter resulta dificil dividirlo en partes iguales. Esto es porque la división está pensada para dividir al texto en secciones que posiblemente hablen de distintos temas y nuestros textos son tweets que por lo general no superan las 10 palabras. Otra dificultad que surge de esta métrica es la imposibilidad de realizar la media de todas las posibles permutaciones del texto por la limitación computacional ya que tenemos una cantidad muy grande de datos.

Es por eso que realizamos una métrica parecida:

Podemos pensar a las palabras del texto como una variable aleatoria W, donde cada palabra w tiene una probabilidad de aparición en una provincia dada de la Argentina.

Esta probabilidad la aproximamos con la frecuencia en la que aparece, es decir la cantidad de ocurrencias de la palabra dividida por la cantidad de palabras totales. Por otro lado sea P una variable aleatoria que cuenta la cantidad de personas que utilizan la palabra p en cada provincia.

Luego,

$$I(w) = I_p(w) * I_u(w) \tag{2.2}$$

$$I_p(w) = norm_p(w) * (\widehat{H}_w(w) - H_w(w)) \tag{2.3}$$

$$I_u(w) = norm_u(w) * (\widehat{H}_u(u) - H_u(w)) \tag{2.4}$$

$$norm_p(p) = \frac{cw(p) - MIN_W}{MAX_W - MIN_W} \tag{2.5}$$

donde: $MIN_W = \min_{p \in Palabras} cw(w)$ (2.6) $MAX_W = \max_{p \in Palabras} cw(w)$ (2.7)

donde $cw(p)$ es igual al logaritmo sobre la cantidad de ocurrencias de esa palabra en toda la Argentina, es decir $cw(p) = \log_2(cantidadOcurrencias(p))$.

Análogamente,

$$norm_u(w) = \frac{cu(w) - MIN_U}{MAX_U - MIN_U} \tag{2.8}$$

$$MIN_U = \min_{p \in Palabras} cu(p) \tag{2.9}$$ $$MAX_U = \max_{p \in Palabras} cu(p) \tag{2.10}$$

donde $cu(p)$ es el logaritmo sobre la cantidad de usuarios que utilizan dicha palabra en la Argentina, es decir $cu(p) = \log_2(cantidadUsuarios(p)))$. $\widehat{H}$ es la entropía con las cantidades distribuidas uniformemente y H es la entropía común.

Tanto $norm_w$ como $norm_u$ realizan una normalización del logaritmo de esas variables. Esto se debe a que el logaritmo genera una dispersión en las medidas de forma tal que su distribución sea más uniforme a lo largo de todo el rango de valores. Esto se puede ver en la figura 2.4. $\widehat{H}_u$ y $\widehat{H}_w$ se corresponde a las entropías de los vectores simulados de apariciones. Esta simulación se realiza con una distribución multinomial ya que se distribuye la suma de los valores de la variable aleatoria uniformemente.

El término de la diferencia de la entropía sobre la cantidad de personas que utilizan la palabra tiene como objetivo mitigar el ruido de la entropía de palabras. En particular una determinada provincia o región pueden tener muchas ocurrencias de una palabra causado por algunos usuarios que utilizan constantemente el término. Un ejemplo de esto podrían ser bots que escriben automáticamente textos iguales (o similares) en grandes cantidades. Otra posible causa de este fenomeno podría ser la de usuarios de ciertas organizacioenes que hablan de personas, lugares o marcas de forma constante. Para eliminar outliers se procedió a eliminar las palabras que tenían una cantidad de usuarios menor o igual a 40 ocurrencias, como también las palabras que eran dichas por menos de 6 personas.

2.5.3. Frecuencia de las palabras

A continuación podemos ver la distribución de la cantidad de ocurrencias de las palabras.

En la figura 2.5 podemos observar que la mayoría de las palabras ocurren poco. En particular el 50 % de las palabras ocurren menos de 139 veces. Por otro lado hay pocas palabras que ocurren mucho, por ejemplo la palabra *que* o la preposición *de*.

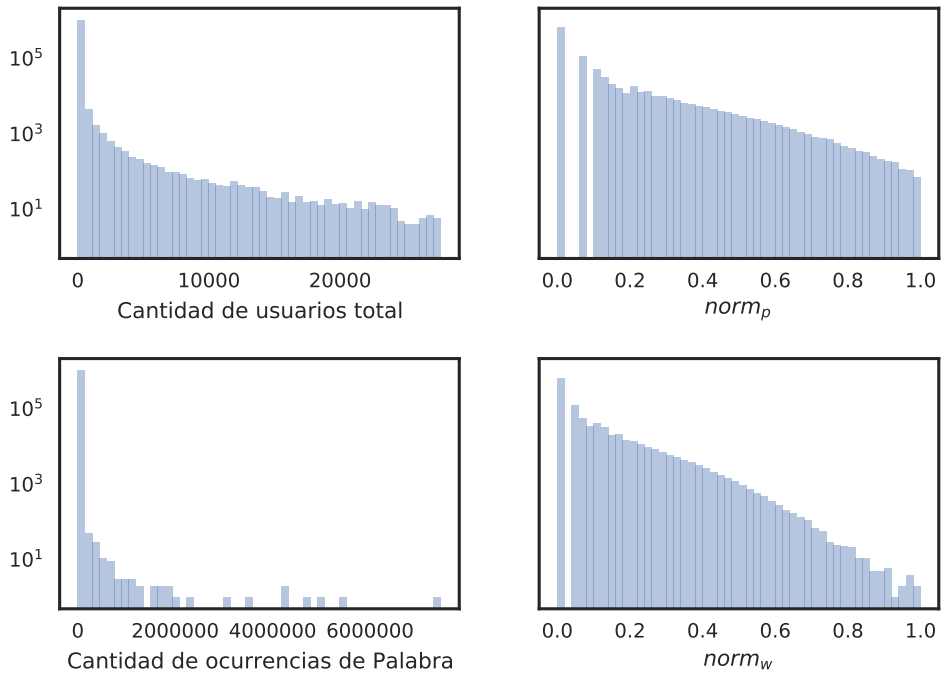


Fig. 2.4: Cantidades y sus normalizaciones

Si comparamos la posición de la palabra en un listado ordenado podemos ver que se cumple con la ley de Zipf. Esta es una ley empírica formulada por George Zipf en el año 1932 en la cual se establece una relación entre la frecuencia de una palabra con su posición dentro del listado de palabras ordenadas por frecuencia decreciente. En particular, sea s la posición de la palabra en el listado ordenado y sea $f(s)$ la cantidad de ocurrencias de la palabra, se puede hacer la siguiente aproximación:

$$f(s) \approx \frac{A}{s^\alpha}$$

donde α toma un valor levemente mayor a 1 y A es una constante.

2.5.4. Test Hipergeométrico

Luego de realizar el listado de palabras ordenado por el valor de la información se realizó un test estadístico para tener mayor confianza de que las palabras clasificadas como contrastivas realmente tienen esta propiedad y no fueron producto del azar. Se seleccionó un conjunto de palabras significativas a nivel lingüistico a partir de las 10000 palabras consideradas más contrastivas por nuestra métrica.

Decidimos elegir el test hipergeométrico ya que querémos ver que la palabra sobre la que se hace el test no estuvo sobrerrepresentada en comparación con la población. Asumimos que la cantidad de ocurrencias de una palabra se puede modelar con una distribución hipergeométrica ya que se puede pensar como un experimento donde se obtuvieron k palabras exitosas en una región con n palabras y un total de N palabras en la Argentina.

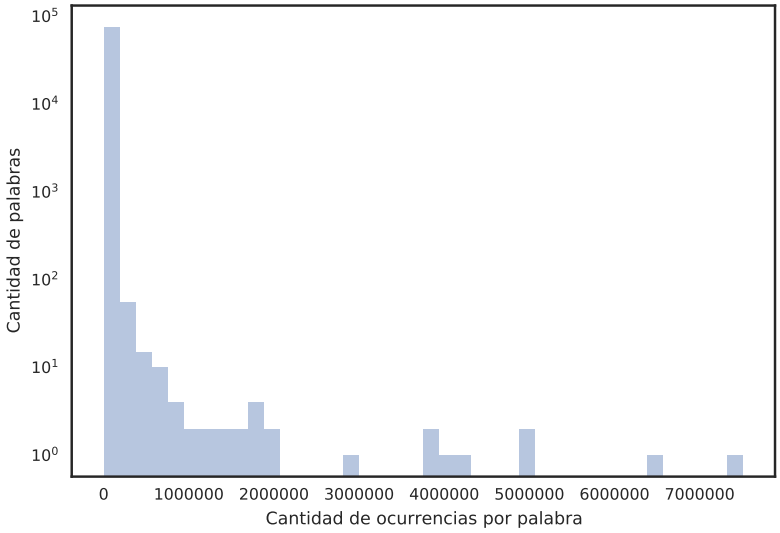


Fig. 2.5: Histograma de la cantidad de ocurrencias de las palabras

Las regiones que utilizamos para cada palabra, son el conjunto de provincias que cubren el 80 % de las ocurrencias de dicho término. Luego, queremos calcular la significancia estadística de haber obtenido esas k palabras exitosas.

La hipótesis nula consiste en que la cantidad de ocurrencias de la palabra en la región elegida es mayor a lo observado. Por lo tanto, sea cantPalabrasW(Region) igual a la cantidad de ocurrencias observada de la palabra en la región a analizar.

$$\begin{cases} H_0 : x > cantPalabrasW(Region) \\ H_1 : x \leq cantPalabrasW(Region) \end{cases}$$

siendo x la esperanza de la variable aleatoria que representa la cantidad de palabras exitosas en esa región.

Palabra	Cantidad de Ocurrencias
que	7509160
de	6527014
a	4962492
la	4913854
no	4177810
me	4101998
y	3838370
el	3773455
en	2969783
te	2060662
se	1976027
un	1863075
es	1825892
con	1799979
lo	1712189
mi	1643777
por	1553382
los	1498941
para	1398757
las	1212452

Tab. 2.2: Cantidad de apariciones de las 20 palabras más frecuentes

	#Palabras Sobre Region	#Palabras en el resto de Argentina	Total
# Palabras w	k	K-k	K
# Palabras ≠ w	n-k	N + k - n - K	N - K
Total	n	N -n	N

Tab. 2.3: Tabla de contingencia

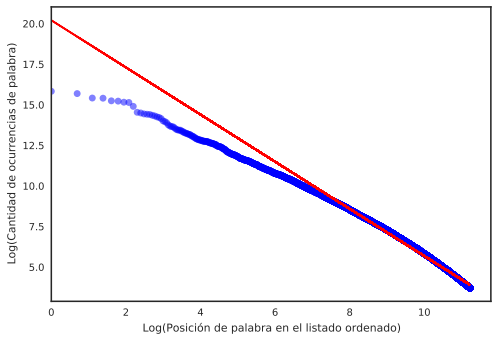


Fig. 2.6: Cantidad de Ocurrencias de palabra vs posición en listado ordenado. Se aplicó el logaritmo a las cantidades de ocurrencias, como también a los valores de las posiciones para mostrar la proporcionalidad entre $f(s)$ y $\frac{1}{s^{\alpha}}$

3. RESULTADOS

3.0.1. Entropía

Teniendo el listado de palabras hicimos un cálculo de entropía tomando en cada provincia la cantidad de ocurrencias de cada palabra. A continuación podemos observar la distribución del valor de la entropía sobre todas las palabras con más de 40 ocurrencias y dichas por más de 5 usuarios:

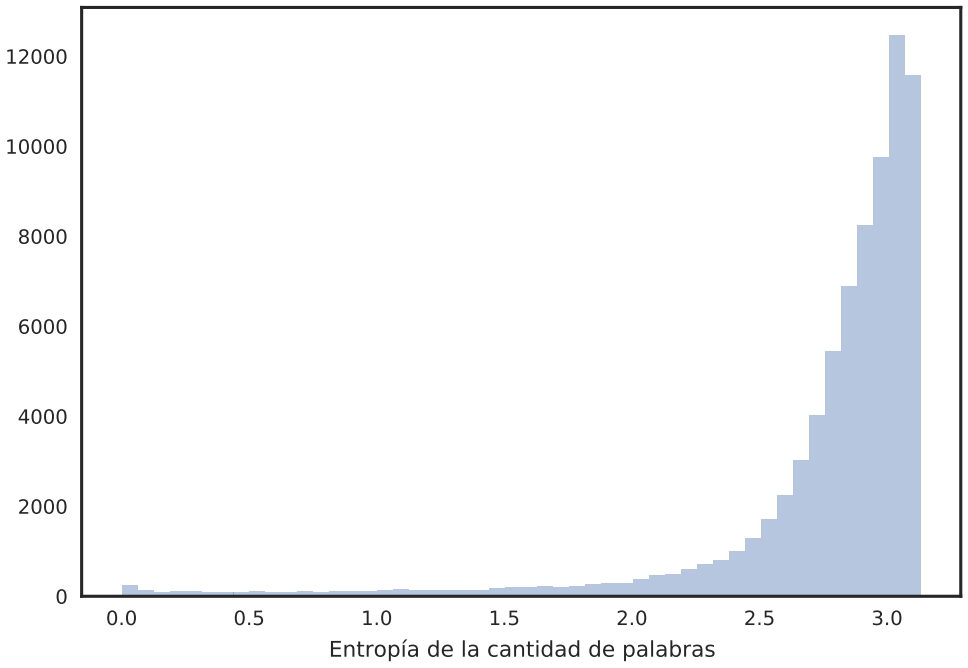


Fig. 3.1: Entropía de las Palabras

En el gráfico 3.1 podemos ver que la mayor parte de las palabras tienen un valor de entropía entre 2.5 y 3. Esto quiere decir que hay un gran conjunto de palabras que tiene una cantidad de ocurrencias relativamente uniforme a lo largo de todas las provincias. Sin embargo hay otro conjunto de palabras que tienen una entropía menor a 2, la cual podemos considerar como baja. Estas útlimas palabras serán las que tienen mayor interés debido a que tienen una variación marcada en cuanto a su utilización en las distintas regiones.

Sin embargo, ver solamente la entropía de las palabras nos puede generar la detección de palabras que no son de interés, ya sea porque no ocurren una cantidad significativa o porque la variación de las ocurrencias en las distintas provincias se debe solamente a pocos usuarios que la utilizan mucho. Es por esto que también se calculó la entropía teniendo como variable la cantidad de personas que utilizaron cierto término en una determinada provincia.

3.0.2. Valor de la información

En el gráfico 3.2 se muestra una clara relación entre la cantidad de ocurrencias que tiene una palabra y su valor de la información, indicado por el color que tiene. A su vez,

se nota que el valor de la información suele ser mayor a médida que el valor de la entropía es menor. Esto no siempre es el caso debido a que hay palabras que tiene una entropía de palabras baja, pero sin embargo la entropía de personas es alta logrando que el valor de la información sea bajo.

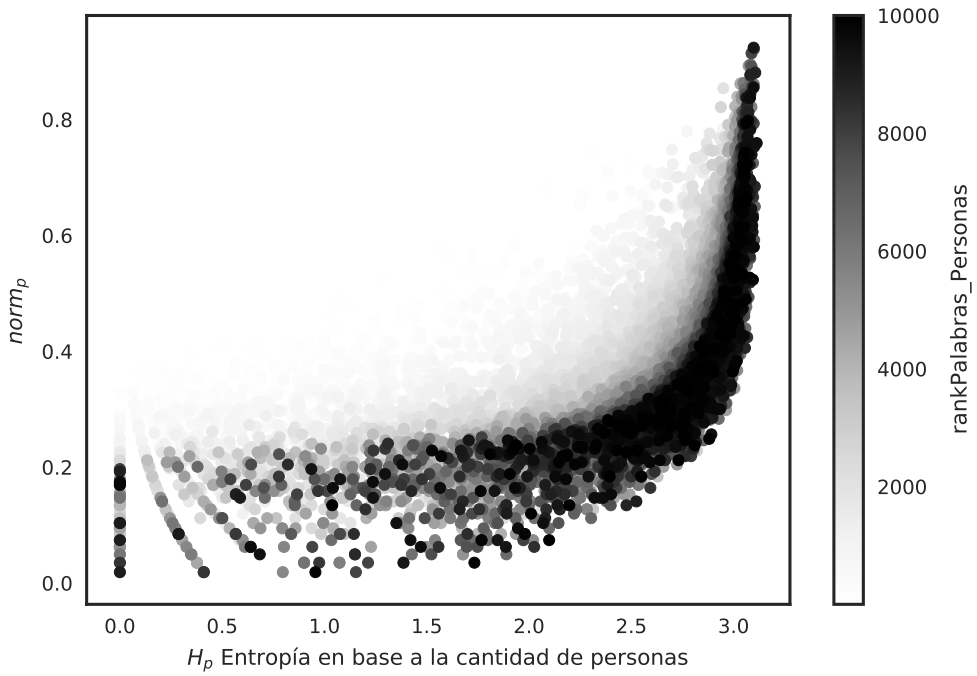


Fig. 3.2: Entropía de las Palabras

A continuación se puede ver el valor de la información según la posición en la que se encuentra en el listado ordenado por la misma.

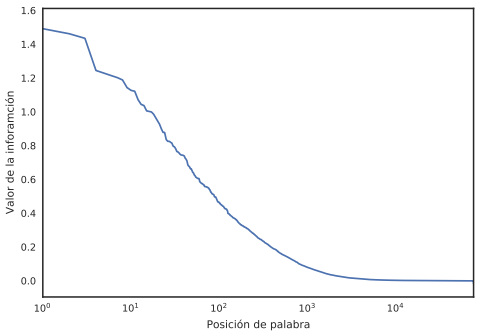


Fig. 3.3: Valor de la información

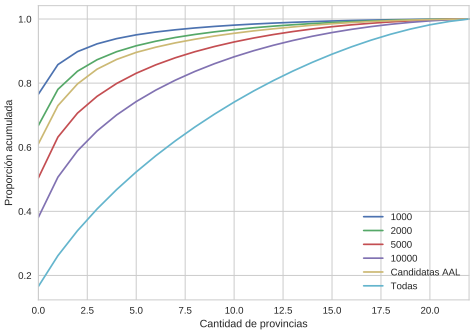


Fig. 3.4: Proporción acumulada según la muestra de palabras

3.0.3. Proporción de ocurrencias

En la figura 3.4 muestra la proporción acumulada de las palabras. Para esto, se ordenó para cada palabra las provincias según la cantidad de ocurrencias. Es notable la diferencia de proporciones acumuladas según la muestra de palabras. Solamente con una provincia para cada palabra ya se puede cubrir, en promedio, el 76 % del total de ocurrencias sobre las mil palabras con mayor valor de la información.

En el gráfico 3.5 se observa la variación del cubrimiento de ocurrencias a médida que se aumenta la cantidad de provincias.

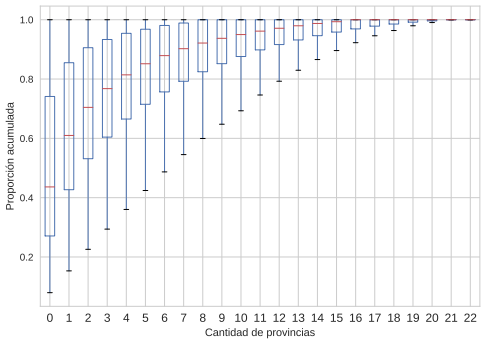


Fig. 3.5: Proporciones Acumuladas

3.0.4. Palabras candidatas

Para buscar las palabras candidatas a tener contrastes significativos en cuanto a la cantidad de ocurrencias en distintas provincias, elegimos el conjunto de las primeras diez mil (10000) palabras con valor de la información más altas. El número 10000 surgió de ver la distribución de los valores de la información. Cómo se puede ver en el gráfico 3.3 hay una caída pronunciada de la métrica y a partir de la palabra cuya posición es 4000 se ve que empieza a establizarse y los valores son muy cercanos a 0. Es por esto que nos pareción razonable dar un márgen de 10000 palabras para seleccionar las palabras con contrastes significativos.

Como era de esperar las ciudades y provincias son palabras que ocurren mayormente en las provinicias que estas indican. Es por esto que tienen una gran variación en la cantidad de ocurrencias en las distintas provincias, lo que genera un valor alto en la métrica de valor de la información. Para detectar las palabras que tienen mayor interés lingüistico buscamnos un conjunto de datos con los nombres de las localidades y departamentos de la República Argentina de modo tal que podamos filtrar los lugares del listado original.

3.0.5. Regiones de palabras

Una vez que calculamos las regiones que cubren un umbral para cada palabra, nos propusimos analizar cuales son las más frecuentes. Para eso generamos una lista con los conjuntos de provincias cuya cantidad sea menor o igual a 6 y los ordenamos según su frecuencia. A continuación se puede ver los primeros conjuntos de provincias obtenidos a partir de los primeras 5000 palabras más contrastivas de acuerdo al valor de la información:

Conjunto de provincias	Cantidad de Palabras
(Jujuy,Salta)	24
(Mendoza, San Juan)	19
(Neuquén, Río Negro)	18
(Corrientes, Misiones)	16
(Chaco, Corrientes, Formosa)	16
(Chaco, Corrientes)	16
(chubut, Santa Cruz)	13
(Catamarca, La Rioja)	12
(Santa Cruz, Tierra del Fuego)	12
(Corrientes, Entre Ríos, Formosa, La Rioja, Misiones)	12
(Formosa, Misiones)	12
(Corrientes, Formosa, Misiones)	12
(Córdoba, La Rioja)	11
(Catamarca, Salta, Santiago del Estero, Tucumán)	11
(Catamarca, Jujuy, La Rioja, Salta, Santiago del Estero, Tucumán)	10
(Chaco, Corrientes, Misiones)	10
(Chaco, Corrientes, Formosa, Misiones)	9
(Catamarca, Santiago del Estero, Tucumán)	9
(Catamarca, Tucumán)	9
(Salta, Tucumán)	8
(Catamarca, Jujuy, Salta, Santiago del Estero, Tucumán)	8
(Neuquén, San Juan)	7
(chubut, Santa Cruz, Tierra del Fuego)	7
(Buenos Aires, lapampa)	7
(Salta, Santiago del Estero, Tucumán)	7
(Buenos Aires, lapampa, Río Negro)	6
(Corrientes, Formosa)	6
(Catamarca, Jujuy, Salta, Tucumán)	6
(Chaco, Corrientes, Entre Ríos, Formosa, Misiones)	6
(Catamarca, Santiago del Estero)	6

Tab. 3.1: Indica cuantas palabras tienen un cubrimiento del 80 % de sus ocurrencias en cada conjunto de provincias a partir de las 5000 palabras más contrastivas (de acuerdo al valor de la información).

Conjunto de Provincias			
Salta-Jujuy	Mendoza-San Juan	Chubut-Santa Cruz-T. Del Fuego	Chaco-Corrientes-Formosa
tribuno	mza	austral	anga
salteño	zonda	chilote	teresss
orán	secamente	calafa	músicas
tartagal	sanjuaninos	chilota	argela
salteña	sanjuanino	palmaso	angaaa
martearena	asar	vueltines	olo
oran	tomba	riviera	cuchale
yuto	queras		corrientesss
purmamarca	traica		angá
yutos	sopaipillas		cts
gorriti	ardente		correntinas
quijano	secamentes		iburrr
tabacal	jáchal		cheraa
desentierro	virreina		cheraá
huaico	tombaaa		bofill
pichanal	mansooo		receppp
diableros	tombino		
bandy	parisi		
aramayo	asadaaa		
ñaño			
colque			
urkupiña			
juy			
guachipas			

Tab. 3.2: Palabras cubiertas sobre el 80 % de las ocurrencias totales por el conjunto de provincias a partir de las 5000 palabras más contrastivas (de acuerdo al valor de la información).

De la tabla 3.0.5 podemos destacar que la mayoría de las regiones son compuestas por provincias contiguas.

3.0.6. Validación

ACA VA LOS RESULTADOS DEL TEST HIPERGEOMETRICO EN EL CONJUNTO DE VALIDACIÓN

3.0.7. Ruido

Cuando vimos las palabras con mayor valor de la información, nos dimos cuenta de que algunas palabras de la provincia de La Rioja eran provenientes de España. Analizando la causa de este ruido, nos dimos cuenta que la API de Twitter no realiza las búsquedas localizadas como uno esperaría. En particular, no solo se fija en los tweets geolocalizados, sino que también hace una búsqueda inversa a través de los nombres de las ciudades que tienen esa coordenada. Específicamente La Rioja es una provincia Argentina, como así también una provincia de España. Es por eso que al hacer búsquedas con las coordenadas de ciudades de La Rioja en Argentina, tuvimos resultados de tweets de España. A pesar

de que los tweets no fueron escritos en Argentina, consideramos que su cantidad no es lo suficientemente grande como para tener resultados incorrectos.

4. CONCLUSIONES Y TRABAJO FUTURO

4.0.1. Conclusiones

En el presente trabajo recolectamos un conjunto de datos de texto de la Argentina a través de la API de Twitter. Este conjunto lo dividimos en dos, un conjunto para desarrollar una métrica que indique el valor contrastivo de una palabra. El segundo conjunto de datos, independiente del primero en cuanto a usuarios y al período temporal de los textos lo utilizamos para hacer un test estadístico para corroborar que las palabras detectadas como contrastivas según nuestra métrica, no estaban sobrerepresentadas en el conjunto de desarrollo.

La métrica que realizamos usaba la entropía para medir la variación de la cantidad de ocurrencias y de la cantidad de usuarios que la utilizaban en las distintas provincias del país. Creamos un listado de palabras ordenado según el valor de la información calculada, y a partir de ella filtramos las palabras de forma manual eliminando los términos que no tengan valor lingüistico, como los nombres propios (como los nombres de personas, o de lugares). Sobre estas palabras realizamos el test estadístico.

4.0.2. Trabajo Futuro

Uno de los desafíos que quedan para hacer es el de clasificar las regiones en clusters, obteniendo así las regiones dialectales. De esta manera se podría ver la vigencia de las regiones descriptas por Vidal de Battini.

El proceso de normalización se podría mejorar para tener una mejor precisión de las palabras utilizadas. También se podría agregar un sistema de reocnocimiento de nombres de entidades para filtrar los nombres propios de manera tal que el listado de palabras contrastivas tengan menos términos sin interés lingüistico.

Por otro lado, este trabajo se podría realizar sobre todo el conjunto de países hispanoparlantes, de modo tal que se puedan hacer comparaciones entre los mismos y comparar las variaciones entre regiones más grandes.

También queda por hacer un análisis sintáctico de las oraciones, y un análisis estadístico de bigramas. Otro desafío es el de analizar los tweets modelados por cadenas de markov (analizando así bigramas y n-gramas), pudiendo generar un bot que cree tweets, este bot podría ser parametrizado de modo tal que genere textos, teniendo en cuenta únicamente los tweets de determinada región.

5. APÉNDICE

5.0.1. La entropía como médida del desorden

Para ver la cantidad de información que nos aporta cada palabra se hará una introducción a la teoría de la información, especificamente los conceptos que introdujo Claude Shannon[Sha01]. Para entender estos conceptos es útil tener una descripción matemática del mecanismo que genera la información. Para eso se define a la *fuente* que emite señales de un alfabeto $S = \{s_1, s_2, \dots\ s_q\}$ de acuerdo a una función de probabilidad fija. Si la fuente emite señáles estadísticamente independientes decimos que es una *fuente de memoria nula* y un símbolo s está completamente determinado por el alfabeto S y las probabilidades: $P(s_1)\, P(s_2)\ \dots\ P(s_q)$

Sea X una variable aleatoria discreta con posibles valores $\{x_1, x_2, \dots\ x_q\}$ y una función de probabilidad P(X), luego: $\mathrm{H}(X) = \mathrm{E}[\mathrm{I}(X)] = \mathrm{E}[-\log(\mathrm{P}(X))]$. donde X es una variable aleatoria con posibles valores $\{x_1, ..., x_n\}$ y P es una función de probabilidad.

Los símbolos con menor probabilidad son los que aportan más información. Esto va de la mano con nuestra intuición ya que si entendemos a los símbolos como palabras de un texto, las palabras más utilizadas como *de* o *que* aportan menos información que la palabra *celular*. Observaciones:

- La entropía es máxima cuando los eventos de X son equiprobables. En este caso, si hay n eventos con una probabilidad de $\frac{1}{n}$ cada uno , el valor de la entropía es de $\log n$.
- La entropía es 0 si y solo sí todas las probabilidades son 0 a excepeción de una con probabilidad igual a la unidad.

Dado que la entropía es máxima cuando los eventos de X son equiprobables, se suele decir que es una médida del desorden.

6. BIBLIOGRAFÍA

Bibliografía

[AV95] Manuel Almeida and Carmelo Vidal. Variación socioestilística del léxico: un estudio contrastivo. *Boletín de filología*, 35(1):Pág–50, 1995.

[Baa01] R Harald Baayen. *Word frequency distributions*, volume 18. Springer Science & Business Media, 2001.

[Esp] Real Academia Española. Banco de datos (corpes xxi)[en línea]. *Corpus del español del siglo XXI (CORPES)*.

[KG03] Adam Kilgarriff and Gregory Grefenstette. Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3):333–347, 2003.

[MH11] Tony McEnery and Andrew Hardie. *Corpus linguistics: Method, theory and practice*. Cambridge University Press, 2011.

[MZ10] Marcelo A Montemurro and Damián H Zanette. Towards the quantification of the semantic information encoded in written language. *Advances in Complex Systems*, 13(02):135–153, 2010.

[Sha01] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.